2015-07-13

Assessing the Validity and Reliability of Dichotomous Test Results Using Item Response Theory on a Group of First Year Engineering Students

Edmund Nevin
Technological University Dublin, edmund.nevin@tudublin.ie

Avril Behan
Technological University Dublin, avril.behan@tudublin.ie

Gavin Duffy
Technological University Dublin, gavin.duffy@tudublin.ie

See next page for additional authors

Assessing the validity and reliability of dichotomous test results using Item Response Theory on a group of first year engineering students

Edmund Nevin
CREATE, Dublin Institute of Technology, Dublin, Ireland
edmund.nevin@dit.ie

Avril Behan
CREATE, Dublin Institute of Technology, Dublin, Ireland
avril.behan@dit.ie

Gavin Duffy
CREATE, Dublin Institute of Technology, Dublin, Ireland
gavin.duffy@dit.ie

Stephanie Farrell
CREATE, Dublin Institute of Technology, Dublin, Ireland
farrell@rowan.edu

Rachel Harding
CREATE, Dublin Institute of Technology, Dublin, Ireland
rachel.harding@student.dit.ie

Robert Howard
CREATE, Dublin Institute of Technology, Dublin, Ireland
robert.howard@dit.ie

Aaron Mac Raighne
CREATE, Dublin Institute of Technology, Dublin, Ireland
aaron.macraighne@dit.ie

Brian Bowe
CREATE, Dublin Institute of Technology, Dublin, Ireland
brian.bowe@dit.ie

***Abstract**: Traditional measurement instruments employed to assess the performance of student's studying on STEM (Science, Technology, Engineering and Mathematics) related programmes typically involve classification based on final scores. The validity and reliability of these instruments and test forms are important considerations when assessing whether a student understands content and if not, where and in what way they are struggling. The aim of this study is to examine, validate and analyse the test results of first-year engineering student's at an Institute of Higher Learning in Ireland who took the Purdue Spatial Visualisation Test of Rotation (PSVT:R). Results obtained were analysed using the RASCH measurement model to see if it could be used to provide an alternative means of measuring student learning and to help identify those who may require extra assistance to overcome academic deficiencies, particularly where spatial skills have been linked to success. Findings may be used to inform on future improvements to teaching approaches and styles.*

***Keywords**: Item Response Theory, Rasch measurement model, spatial visualisation, PSVT:R*

Introduction

The primary aim of a measurement instrument is to quantify some phenomenon through the assignment of numbers to observations. Two key indicators of the quality of a measuring instrument are its validity and reliability (Kimberlin and Winterstein, 2008).

Multiple choice questions (MCQs) are a popular and widely used instrument for assessing student learning (Huntley et al., 2009; Pande et al., 2013). Different approaches to analysing MCQs exist (Ding and Beichner, 2009), with two popular frameworks being Classical Test Theory (CTT) and Item Response Theory (IRT) (Hambleton and Jones, 1993; Thorpe and Favia, 2012). Both attempt to align test-takers on a scale or latent trait continuum. The latent variable is typically a hypothetical construct such as ability, which is suggested to exist but cannot be measured by a single observable variable or item. Measurement of the latent variable is carried out indirectly through a test instrument consisting of multiple items.

CTT is a term which encompasses several types of psychometric tests. Most approaches assume that the observed score (X) obtained by the test-taker is made up of a true score (T) and a random error (E) giving $X = T + E$. IRT, on the other hand, takes what is known about the items e.g. difficulty, discrimination and the pattern of responses to the item and then makes an estimate of a person's most likely level of the trait being measured e.g. ability. According to Fayers and Hayes (2005, p. 55), "*IRT refers to a set of mathematical models that describe, in probabilistic terms, the relationship between a person's response to a survey question/test item and his or her level of the* '*latent variable*' *being measured by the scale*".

The Rasch measurement model was chosen as it is widely recognised as being a robust and objective measurement of latent traits (Hendriks et al., 2012). Its application can be found across many disciplines including, but not limited to, health, social sciences and education (Bonsaksen et al., 2013; Hudson and Treagust, 2013; Lerdal et al., 2014). A number of key assumptions underpin the Rasch model (Fischer, 1974).

1. *Unidimensionality*: All items are functionally dependent upon only one underlying continuum i.e. only one underlying factor accounts for a person's response to a question within a scale.
2. *Monotonicity*: All item characteristic functions (ICF) are strictly monotonic in the latent trait. The ICF describes the probability of a predefined response as a function of the latent trait.
3. *Local stochastic independence*: Every person has a certain probability of giving a predefined response to each item and this probability is independent to the answers given to the preceding items.
4. *Sufficiency of a simple sum statistic*: The number of predefined responses is a sufficient statistic for the latent parameter.
5. *Dichotomy of the items*: For each item there are only two different responses such as yes/no, true/false, or agree/disagree.

Aim of Study

Research studies often present data obtained from test instruments without a rigorous critical reflection on what the data obtained from the instruments actually means despite numerous statistical tests existing to measure the validity and reliability of test instruments. The primary aim of this study is to examine how IRT can be used to determine the validity and reliability of data obtained. For this reason a popular test instrument (PSVT:R) used in engineering and other STEM related disciplines to evaluate the spatial ability of test-takers was chosen. While numerous studies have utilised spatial visualisation tests to measure the spatial ability of students' (Sorby and Baartmans, 2000; Towle et al., 2005; Hamlin et al., 2006), less attention

has been given to examining the validity and reliability of the instrument measure with one notable exception being Maeda and Yoon (2011).

Validity and Reliability

Validity and reliability are key concepts in measurement. In order to be useful, measurement instruments should be both valid and reliable. According to Messick (1993) "*Validity is an overall evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores*". Reliability is the extent to which the instrument consistently measures what it is intended to measure. While validity requires an instrument to be reliable, the reverse doesn't hold as an instrument can be reliable without being valid (Kimberlin and Winterstein, 2008).

Formal definitions of validity vary so in an attempt to unify the theory of construct validity, in other words, how well the instrument does what it claims to do, Messick (1995) proposed a six faceted construct to measure the quality of the instrument: content, substantive, structural, generalisability, external and consequential. These facets have been used in numerous studies to validate the outputs produced by Rasch models (Wolfe and Smith, 2007; Beglar, 2010; Baghaei and Amrahi, 2011).

Rasch Measurement Model

The basis of the Rasch measurement model is that for each person taking a test there is an unobservable construct (or latent variable) being measured by a scale i.e. ability (θ) and for each item on the test there is a parameter that measures the difficulty of an item response (β). Using these parameters Rasch (1960) proposed that the level of learning may be determined through the interaction $\theta_i - \beta_j$ where θ_i is the score of the i-th student (i.e. ability) and β_j is the score of the j-th item (i.e. difficulty). The probabilistic model for dichotomous data is given as:

$$p(x_{ij} = 1|\theta_i, \beta_j) = \frac{e^{(\theta_i - \beta_j)}}{1 + e^{(\theta_i - \beta_j)}} = p_{ij} \qquad \text{(i)}$$

Equation (i) states how likely a person is to endorse a response category depends on how much of the trait they have and how difficult the item is. Data is collected and stored in a matrix form as depicted in Table 1. The table is made up of one row for each person (i.e. n rows) and one column for each item (i.e. j columns). Correct answers are indicated with '1' and incorrect with '0'. The total score of student i for all items is given by the sum of each row i.e. $r_i = \sum_{j=1}^{n} x_{ij}$. The score given by all students to item j is the sum of each column i.e. $s_i = \sum_{i=1}^{n} x_{ij}$.

Due to the non-linear nature of the scores, a direct comparison between row and column totals is not possible. Rasch analysis converts the raw scores into linear units of measure called 'logits'. The Rasch model uses a logit scale for both θ_i and β_j. The logit of p represents the log-odds of correctly answering an item. By taking the natural logarithm of both sides of equation (i), equation (ii) is obtained.

$$\log_e\left(\frac{p}{1-p}\right) = \theta_i - \beta_j \qquad \text{(ii)}$$

i.e.
$$\log\left(\frac{probability\ of\ success}{probability\ of\ failure}\right) = Ability - Difficulty \qquad \text{(iii)}$$

where
$$p = p(x_{ij} = 1|\theta_i, \beta_j) \qquad \text{(iv)}$$

Using logits makes it easier to make direct comparisons between student ability (θ_i) and item difficulty (β_j). The proportion correct is simply an average of the column (item) or row (person) scores. Ability levels are obtained by taking the natural log of a ratio of the proportion correct to the proportion incorrect as illustrated by equation (v). Item difficulty is

obtained by taking the natural log of a ratio of the proportion incorrect responses to the proportion correct as illustrated by equation (vi).

Table 1: Determining item difficulty and ability estimates.

Person	Item 1	Item 2	Item 3	Item 4	Item 5	Proportion Correct (P)	Ability (θ)
1	1	0	0	0	0	0.20	-1.39
2	1	1	0	1	0	0.60	0.41
3	1	1	1	0	0	0.60	0.41
4	1	1	0	1	0	0.60	0.41
5	1	1	1	0	1	0.80	1.39
6	0	1	1	0	0	0.20	-1.39

Proportion Correct (P)	0.83	0.67	0.50	0.33	0.17
Item Difficulty (β)	-1.61	-0.69	0.00	0.69	1.61

$$\theta_4 = \ln\left(\frac{P_4}{1-P_4}\right) = \ln\left(\frac{0.6}{1-0.6}\right) = 0.41 \quad \text{(v)}$$

$$\beta_5 = \ln\left(\frac{1-\beta_5}{\beta_5}\right) = \ln\left(\frac{0.17}{1-0.17}\right) = 1.61 \quad \text{(vi)}$$

The conceptualisation of the ability (latent) continuum as a ruler is illustrated in Figure 1. Person ability and item difficulty are converted into linear interval measures using a log-odds (logit) transformation. The mean item difficulty is assigned a logit value of 0 as the difference between person ability and item difficulty is not absolute but relative. The trait being measured may now be determined on a linear scale.

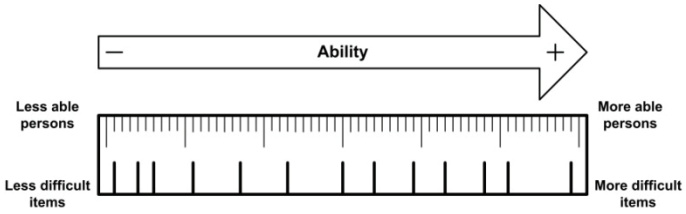

Figure 1: Conceptualisation of the ability continuum.

According to Harris (1989) item difficulty values can theoretically range from $-\infty \leq \beta \leq +\infty$ but in reality generally range between $-3 \leq \beta \leq +3$ with values in excess of ± 3 rare. Similarly, person ability values can theoretically range from $-\infty \leq \theta \leq +\infty$ but will generally lie in the range $-3 \leq \theta \leq +3$.

Methodology

Participants

Test results from a sample of 236 students, who took the PSVT:R, form the basis of the analysis for this study. The test was administered to a cohort of first-year engineering (FYE) students at Dublin Institute of Technology at the start of their first semester (2014-2015).

Instrument

The PSVT:R is a widely accepted and respected instrument used in engineering education to measure the 3D visualisation ability of students'. It was developed by Guay (1976) at Purdue University and consists of 30 items drawn in 2D isometric format. An example problem from the PSVT:R is shown in Figure 2. Here an object is shown on the top line which has been rotated by a given amount. Below this, a second object is shown which the test-taker must mentally rotate by the same amount and the correct view must then be chosen from the third line. Each item has only one correct answer. Students have 20 minutes to complete the test.

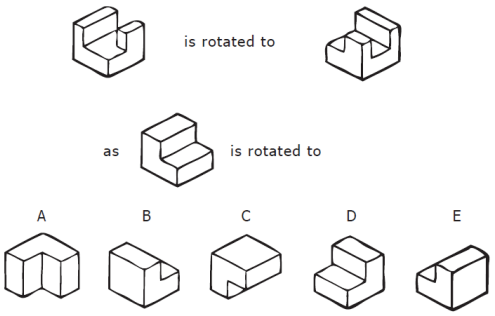

Figure 2: Example problem from PSVT:R (correct answer = D)

Statistical Analysis

All test data was analysed using Winsteps® Rasch Measurement software version 3.81.0 (Linacre, 2014) with results presented in Table 4. The sequence of steps outlined are based on a number of the construct validity facets proposed by Messick (1995), while person and item reliability are measured using the person separation index and Cronbach's alpha coefficient (Cronbach, 1951). Unlike CTT, Rasch measurement does not require complete data sets so incomplete responses do not adversely affect the analysis.

Results

Summary Statistics

Table 2 provides summary statistics for both PERSON (test-takers) and ITEM (quiz question) measures. The value for MEASURE represents the estimated measure (for persons) or calibration (for items). In the Rasch measurement model INFIT is an inlier-pattern-sensitive fit statistic based on the chi-square (mean-square) statistic. It is more sensitive to unexpected behaviour affecting responses near to the measure level. OUTFIT is an outlier-sensitive fit statistic based on the chi-square statistic and is more sensitive to unexpected observations by persons on items that are relatively very easy or very hard for them (and vice-versa).

Table 2: Summary statistics from Winsteps® for both PERSON and ITEM measures.

PERSON	236 INPUT	236 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	20.7	30.0	1.17	.55	.99	.0	1.03	.1
S.D.	6.0	.0	1.37	.24	.20	1.0	.50	1.1
REAL RMSE	.60	TRUE SD	1.23	SEPARATION	2.05	PERSON RELIABILITY		.81
ITEM	30 INPUT	30 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	162.9	236.0	.00	.17	.99	-.1	1.03	.2
S.D.	34.4	.0	.92	.02	.07	.9	.14	.9
REAL RMSE	.17	TRUE SD	.91	SEPARATION	5.18	ITEM	RELIABILITY	.96

Fit statistics in the form of mean square (MNSQ) and standardised fit (ZSTD) are used to determine the goodness-of-fit of both PERSON and ITEM measures. MNSQ is a chi-square statistic used to compare expected results with those actually observed. Its value should be close to 1. A MNSQ value >1 indicates underfit (existence of embedded noise) while a value < 1 indicates overfit (results in inflated summary statistics). ZSTD reports the statistical significance (probability) of the chi-square statistics occurring by chance when the data fits the Rasch model i.e. it indicates how accurately or predictably data fits the model. The expected value for ZSTD is 0. A value <0 indicates that the data is too predictable while a >0 indicates a lack of predictability.

Item Difficulty

Item difficulty is estimated by the Rasch model. Table 3 provides a summary of the items on the PSVT:R in descending order of difficulty as determined by the Rasch model. Here 'measure' refers to the item's measure calibration i.e. the higher the value the more difficult the test item. 'Rotation' describes the number of axis rotations required for an item and

whether the rotation is symmetrical (*S*) i.e. rotation confined to one axis or non-symmetrical (*NS*) i.e. multiple axis-rotations are required to arrive at the solution.

Table 3: PSVT:R item difficulty as determined by the Rasch model.

Difficulty	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Item #	Q30	Q29	Q27	Q26	Q22	Q28	Q13	Q25	Q12	Q23	Q19	Q20	Q14	Q21	Q17
Rotation	3-*NS*	3-*NS*	3-*NS*	3-*NS*	2-*NS*	3-*NS*	2-*S*	3-*NS*	2-*S*	3-*NS*	2-*NS*	2-*NS*	2-*S*	2-*NS*	2-*NS*
Measure	2.74	1.44	1.25	1.16	1.00	0.96	0.56	0.56	0.25	0.25	0.20	0.13	-0.03	-0.05	-0.08
Difficulty	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
Item #	Q24	Q15	Q18	Q16	Q10	Q11	Q6	Q7	Q8	Q3	Q5	Q2	Q9	Q1	Q4
Rotation	3-*NS*	2-*NS*	2-*NS*	2-*NS*	2-*S*	2-*S*	1-*S*	2-*S*	2-*S*	1-*S*	1-*S*	1-*S*	2-*S*	1-*S*	1-*S*
Measure	-0.08	-0.11	-0.33	-0.36	-0.39	-0.42	-0.48	-0.51	-0.74	-0.81	-0.99	-1.03	-1.15	-1.19	-1.79

Table 4: Summary of validity and reliability checks obtained from the Rasch model.

Step	**Psychometric Property**	**Statistical Approach and Criteria**	**Results**
1	**Rating scale functioning:** (substantive validity) Does the rating scale function consistently across items?	• Average measures for each step category and threshold on each item should advance monotonically. • z-values <2.0 in OUTFIT mean square (MNSQ) values for step category calculations.[a]	• Item 13 did not meet criteria (z-value = 2.1) • Item 14 did not meet criteria (z-value = 2.0)
2	**Internal scale validity:** (content validity) How well do the actual item responses match the expected responses from the Rasch model?	**Item goodness-of-fit statistics:** • MNSQ <1.3[b]	• All items met criterion
3	**Internal scale validity:** (structural validity) Is the scale unidimensional?	**Principal component analysis:** • ≥50% of total variance explained by first component (spatial ability).[c] • Any additional component explains <5% (or Eigenvalue <2.0) of the remaining variance after removing the first component.[c] • No more than 5% (1 out of 20) of the residual correlations >0.30	• First component explained 70.0% of total variance. • Second component explained 5.2% of total variance with an eigenvalue = 2.2 (>2.0). • Two out of 20 residual correlations >0.3 (item 1 - item 2: r = 0.57, item 3 - item 4: r = 0.35)
4	**Person response validity:** (substantive validity) How well do the actual individual responses match the expected responses from the Rasch model?	**Person goodness-of-fit statistics:** • INFIT MNSQ values <1.5 and z-value ≤2.0.[d] • ≤5% of sample fails to demonstrate acceptable goodness-of-fit values.[d]	• 1 out of 236 (<< 5%) respondents failed to demonstrate acceptable goodness-of-fit values
5	**Person Separation Reliability:** (reliability) Can the scale distinguish at least two distinct levels of sense of coherence in the sample tested?	**Person Separation index:** • ≥2.0[e]	• 2.07 (Real i.e. lower bound) • 2.17 (Model i.e. upper bound)
6	**Internal Consistency:** (reliability) Are item responses consistent with each other?	**Cronbach's alpha coefficient:** • ≥0.8[e]	• 0.87

[a] Linacre (2002)
[b] Wright et al. (1994)
[c] Linacre (2014)
[d] Kottorp et al. (2003)
[e] Fisher (1992)

Validity and Reliability

A summary of the validity (steps 1-4) and reliability (steps 5-6) checks are given in Table 4, the layout of which is adopted from Bonsaksen et al. (2013). The statistical approach and criterion for each of the psychometric properties investigated are outlined. The validity checks incorporate three (substantive, content and structural) of the six distinguishable aspects of unified validity put forward by Messick (1995). Two reliability coefficients are used i.e. the separation and reliability indices. The separation index represents how well the measurement instrument can distinguish between persons based on their ability location. Values between 2 and 3 for the separation index are considered to be very good levels for separation capacity (Fisher, 1992). The reliability index used in Rasch analysis is similar to Cronbach's alpha (Bond and Fox, 2007).

Person Responses

Once the reliability and validity of the test instrument has been established the Rasch model can be used to examine the individual responses of the test-takers. In one example, Edwards and Alcock (2010) use the Rasch model to examine uncharacteristic responses.

For this study, Table 5 lists the 14 test-takers whose responses most misfit the Rasch model i.e. their responses differ from those estimated by the Rasch model. Fit statistics based on MNSQ and ZSTD values were used to identify test-takers who did not fit the Rasch model. The acceptable range of MNSQ is from 0.8 to 1.2 (Wright 1994) and ZSTD values are between -2 and 2 (Bond and Fox, 2007).

With reference to Table 5, consider the following:

- A large outfit-ZSTD value (>2) coupled with a high measure may indicate that a student has answered an 'easy' question incorrectly. In this study person 033 (score = 93%) answered item Q2 (measure = -1.03) incorrectly.
- A large outfit-ZSTD value (>2) coupled with a low measure may indicate that a student has answered a 'tough' question correctly and the remainder mostly incorrectly. In this study person 841 (score = 50%) answered item Q2 (difficulty = 2.74) correctly.

Table 5: Output from the Rasch model identifying misfit respondents.

#	Person	Total Score (/30)	Measure	OUTFIT		#	Person	Total Score (/30)	Measure	OUTFIT	
				ZSTD	MNSQ					ZSTD	MNSQ
1.	841	15	-0.03	5.2724	2.3986	**8.**	596	24	1.6	2.1821	2.1409
2.	659	16	0.13	3.1517	1.7253	**9.**	525	18	0.45	2.0215	1.4671
3.	536	16	0.13	3.1217	1.7157	**10.**	923	22	1.16	-2.0095	0.464
4.	973	27	2.53	2.5037	3.7344	**11.**	668	20	0.79	-2.0394	0.554
5.	677	17	0.28	2.3615	1.5263	**12.**	212	17	0.28	-2.0794	0.6316
6.	017	25	1.86	2.2925	2.4689	**13.**	564	18	0.45	-2.1094	0.6082
7.	033	28	3.03	2.2743	4.2725	**14.**	921	17	0.28	-2.2194	0.6119

Discussion and Conclusions

Research has shown that results from the PSVT:R may be used as key indicators of success in STEM related disciplines (Humphreys, Lubinski, & Yao, 1993; Sorby, 2009; Wai, Lubinski, Benbow, & Steiger, 2010). For this study, the validity and reliability of the test instrument was reinforced by the results obtained and documented in Table 4. Overall, both person and item measures demonstrated acceptable goodness-of-fit and are positive indicators to the quality of the data and the test instrument.

While two items did not meet the criteria set out in Step 1 (see Table 4), they were only marginally outside the range and were not excluded in this instance. In terms of variance (see Table 6), the Rasch model explained 70.0% of the total variance in the dataset (i.e.

spatial ability). The secondary dimension explained 5.2%. As the first contrast is not much larger than the size of the eigenvalue expected by chance (<2.0) and although two out of 20 residual correlations were found to be >0.3 (see Table 4) there is strong enough evidence to support the existence of local independence of the items i.e. unidimensionality.

Table 6: Output from Winsteps® showing standardised residual variance values.

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)

		-- Observed --		Expected
Total raw variance in observations	=	42.9	100.0%	100.0%
Raw variance explained by measures	=	12.9	30.0%	29.6%
Raw variance explained by persons	=	6.4	15.0%	14.7%
Raw Variance explained by items	=	6.4	15.0%	14.8%
Raw unexplained variance (total)	=	30.0	70.0% 100.0%	70.4%
Unexplned variance in 1st contrast	=	2.2	5.2% 7.4%	
Unexplned variance in 2nd contrast	=	1.8	4.3% 6.1%	
Unexplned variance in 3rd contrast	=	1.7	3.9% 5.6%	
Unexplned variance in 4th contrast	=	1.6	3.6% 5.2%	
Unexplned variance in 5th contrast	=	1.5	3.5% 4.9%	

A high positive residual correlation (r) can indicate local item dependency whereas a large negative correlation indicates the opposite of local dependence. The residual correlation obtained from this study was r = 0.4 which may be considered low dependency. With a person-separation index of 2.07 combined with a reliability of 0.87 (see Table 4) it can be concluded that the test instrument used was able to distinguish between two categories of test-takers. A value of 2.07 indicates that the test instrument detected two statistically distinct groups of participants for the trait being investigated i.e. in this case, test-takers with high and low spatial ability. Cronbach's alpha reports the approximate test reliability based on raw scores and with a value of 0.87 obtained from this study is above the acceptable value of 0.8.

The Rasch model has a role to play in both engineering education for assessing students through MCQs, surveys etc., and in engineering education research as a tool for examining the validity and reliability of measures obtained from various test instruments, not just the PSVT:R which was used in this study. As the example provided in this paper illustrates, the RASCH model may be used to provide an alternative means for measuring student learning ability and can help identify those who may require targeted intervention. Findings from this and similar studies may be used to inform on future improvements to teaching approaches and styles.

References

Baghaei, P., & Amrahi, N. (2011). Validation of a Multiple Choice English Vocabulary Test with the Rasch Model. *Journal of Language Teaching and Research*, *2*(5), 1052-1060.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, 27, 101-118.

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch Model*. Routledge.

Bonsaksen, T., Kottorp, A., Gay, C., Fagermoen, M., & Lerdal, A. (2013). Rasch analysis of the General Self-Efficacy Scale in a sample of persons with morbid obesity. *Health and Quality of Life Outcomes*, *11*(1), 1-11.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

Ding, L., & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. Physical Review Special Topics-Physics Education Research, 5(2), 020103.

Edwards, A., & Alcock, L. (2010). Using Rasch analysis to identify uncharacteristic responses to undergraduate assessments. *Teaching Mathematics and its Applications*, *29*(4), 165-175.

Fayers, P., & Hays, R. (2005). *Assessing quality of life in clinical trials* (2nd ed.). OUP Oxford.

Fischer, G. H. (1974) Derivations of the Rasch Model. In Fischer, G. H. & Molenaar, I. W. (Eds) Rasch Models: foundations, recent developments and applications, pp. 15-38 New York: Springer Verlag.

Fisher W (1992). Reliability, separation, strata statistics. *Rasch Measurement Transaction*, 6:238.

Guay, R. B., (1977), Purdue spatial visualization test: Rotations, West Lafayette, IN, Purdue Research Found.

Hambleton, R. K., & Jones, R. W. (1993). An NCME Instructional Module on *Educational Measurement: Issues and Practice*, *12*(3), 38-47.

Hamlin, A., Boersma, N., & Sorby, S. (2006). Do spatial abilities impact the learning of 3D solid modelling software? *Proceedings of the American Society for Engineering Education Conference and Exposition*.

Harris, D. (1989). Comparison of 1-, 2-, and 3-Parameter IRT Models. *Educational Measurement: Issues and Practice*, *8*(1), 35-41.
Hendriks, J., Fyfe, S., Styles, I., Rachel Skinner, S. & Merriman, G. (2012. Scale construction utilising the Rasch unidimensional measurement model: A measurement of adolescent attitudes towards abortion. Australasian Medical Journal, 5(5), 251-261.
Hudson, R. D., & Treagust, D. F. (2013). Which form of assessment provides the best information about student performance in chemistry examinations? *Research in Science & Technological Education*, *31*(1), 49-65.
Huntley, B., Engelbrecht, J., & Harding, A. (2009). Can multiple choice questions be successfully used as an assessment format in undergraduate mathematics? *Pythagoras*, (69), 3-16.
Humphreys, L. G., Lubinski, D., & Yao, G. (1993). Utility of predicting group membership and the role of spatial visualization in becoming an engineer, physical scientist, or artist. *Journal of Applied Psychology, 78*(2), 250-261.
Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *Am J Health Syst Pharm*, *65*(23), 2276-84.
Kottorp, A., Bernspång, B., & Fisher, A. (2003). Validity of a performance assessment of activities of daily living for people with developmental disabilities. *Journal of Intellectual Disability Research*, *47*(8), 597-605.
Lerdal et al. (2014) Rasch analysis of the sense of coherence scale in a sample of people with morbid obesity - a cross-sectional study. *BMC Psychology*, *2*:1.
Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *J Appl Meas*, *3*(1), 85-106.
Linacre, J. M. (2014). A User's Guide to Winsteps/Ministeps Rasch-Model Computer Programs.
Maeda, Y., & Yoon, S. Y. (2011). Scaling the Revised PSVT-R: Characteristics of the first year engineering students' spatial ability. In *Proceedings of the American Society for Engineering Education* (ASEE) Annual Conference and Exposition, 2011-2582, Vancouver, BC, Canada.
Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment*, *1*(1), 1-11.
Messick, S., (1993), *Validity*, in Linn, R.L. Educational Measurement (3rd ed.), Phoenix, AZ: American Council on Education and the Oryx Press.
Messick, S., (1995), Standards of Validity and the Validity of Standards in Performance Assessment. *Educational Measurement: Issues and Practice*, *14*(4), 5-8.
Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danish Institute for Educational Research.
Pande, S. S., Pande, S. R., Parate, V. R., Nikam, A. P., & Agrekar, S. H. (2013). Correlation between difficulty and discrimination indices of MCQs in formative exam in Physiology. *South East Asian Journal of Medical Education*, 7, 45-50.
Sorby, S. A. (2009). Educational Research in Developing 3-D Spatial Skills for Engineering Students. *International Journal of Science Education, 31*(3), 459-480.
Sorby, S. A., & Baartmans, B. J. (2000). The development and assessment of a course for enhancing the 3D spatial visualization skills of first year engineering students. *Journal of Engineering Education, 60*, 301307.
Thorpe, G. L. and Favia, Andrej (2012). Data Analysis Using Item Response Theory Methodology: An Introduction to Selected Programs and Applications. *Psychology Faculty Scholarship*. Paper 20.
Towle, E., Mann, J., Kinsey, B., O'Brien, E. J., Bauer, C. F., & Champoux, R. (2005). Assessing the self-efficacy and spatial ability of engineering students from multiple disciplines. In Frontiers in Education, 2005. *FIE'05. Proceedings 35th Annual Conference* (pp. S2C-15). IEEE.
Wolfe, E. W., & Smith, E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II-validation activities. *J. of App. Measurement*, *8*(2), 204-234.
Wai, J., Lubinski, D., Benbow, C. P., & Steiger, J. H. (2010). Accomplishment in science, technology, engineering, and mathematics (STEM) and its relation to STEM educational dose: A 25-year longitudinal study. Journal of Educational Psychology, 102(4).
Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Copyright statement